CHAPTER-1

INTRODUCTION

1.1 Natural Language Processing:

Natural language processing (NLP) is the ability of a computer program to understand human speech as it is spoken. NLP is a component of artificial intelligence (AI). The development of NLP applications is challenging because computers traditionally require humans to “speak” to them in a programming language that is precise, unambiguous and highly structured or, perhaps through a limited number of clearly-enunciated voice commands. Human speech, however, is not always precise – it is often ambiguous and the linguistic structure can depend on many complex variables, including slang, regionaldialects and social context. Natural language processing (NLP) is a field of computer science, artificial intelligence, and linguistics concerned with the interactions between computers and human (natural) languages. As such, NLP is related to the area of human–computer interaction. Many challenges in NLP involve natural language understanding, that is, enabling computers to derive meaning from human or natural language input, and others involve natural language generation.

1.1.1 History[1]:

The history of NLP generally starts in the 1950s, although work can be found from earlier periods. In 1950, Alan Turing published an article titled ”Computing Machinery and Intelligence” which proposed what is now called the Turing test as a criterion of intelligence. Some notably successful NLP systems developed in the 1960s were SHRDLU, a natural language system working in restricted ”blocks worlds” with restricted vocabularies, and ELIZA, a simulation of a Rogerian psychotherapist, written by Joseph Weizenbaum between 1964 to 1966. Using almost no information about human thought or emotion, ELIZA sometimes provided a startlingly human-like interaction. During the 1970s many programmers began to write ’conceptual ontologies’, which structured real-world information into computer-understandable data. Examples are MARGIE (Schank, 1975), SAM (Cullingford, 1978), PAM (Wilensky, 1978), TaleSpin (Meehan, 1976), QUALM (Lehnert, 1977), Politics (Carbonell, 1979), and Plot Units (Lehnert 1981). During this time, many chatterbots were written including PARRY, Racter, and Jabberwacky. Up to the 1980s, most NLP systems were based on complex sets of hand-written rules. Starting in the late 1980s, however, there was a revolution in NLP with the introduction of machine learning algorithms for language processing. Recent research has increasingly focused on unsupervised and semi-supervised learning algorithms. Such algorithms are able to learn from data that has not been hand-annotated with the desired answers, or using a combination of annotated and non-annotated data.

1.1.2 Major tasks in NLP:

The following is a list of some of the most commonly researched tasks in NLP.

- Automatic summarization
- Coreference resolution
- Discourse analysis
- Machine translation
- Natural language generation
- Part-of-speech tagging
- Speech recognition
- Speech processing

1.2 Software Description:

Our overall goal is to introduce Punjabi language of speech recognition research by releasing state of the art open source speech to text technology, and making massive amounts of speech data freely available. Currently, speech to text technology is only available from a handful languages. We believe that is a barrier to Punjabi language in this market.

Some of the software that are capable of converting speech to text conversion are as follows:

- EvalDictator
- Simon
- Dragon NaturallySpeaking
- Julius

1.2.1 EvalDictator:-

In order to achieve these ends, it popularize speech recognition technology by building open source applications. These tools will be written in Java and will run on every major platform including Windows, OSX and Linux. Its target is computer users who wish to enter text in their native language, and prefer speech to the keyboard. These users may be professionals who require hands free text entry. For example, many Doctors prefer to enter reports via dictation. Another target is users who find it difficult to type text in their native language. For example, in many countries users must memorize many complicated key combination to enter basic text. Large countries (e.g. China, Japan...) and countless smaller ethnic groups have this problem.

Simon:

Simon is an open source speech recognition program that can replace your mouse and keyboard. The system is designed to be as flexible as possible and will work with any language or dialect.

Simon uses the KDE libraries, CMU SPHINX and / or Julius coupled with the HTK and runs on Windows and Linux.

Dragon NaturallySpeaking:

Dragon NaturallySpeaking (also known as Dragon for PC, or DNS), is a speech recognition software package developed by Dragon Systems of Newton, Massachusetts, and later acquired by Nuance Communications. It runs on Windows personal computers.

Julius:

Julius is an open source speech recognition engine.

Julius is a high-performance, two-pass large vocabulary continuous speech recognition (LVCSR) decoder software for speech-related researchers and developers. It can perform almost real-time decoding on most current PCs in 60k word dictation task using word 3-gram and context-dependent HMM.

CHAPTER-2

LITERATURE SURVEY

2.1 Shikano Kiyohiro and Kawahara Tatsuya N[2] in the paper "IPA Japanese Dictation Free Software Project" proposed that Large vocabulary continuous speech recognition (LVCSR) is an importantbasis for the application development of speech recognition technology. We had constructed Japanese common LVCSR speech database and have been developing sharable Japanese LVCSR programs/models by the volunteer based efforts. We have been engaged in the following two volunteer-based activities. a) IPSJ (Information Processing Society of Japan) LVCSR speech database working group. b) IPA (Information Technology Promotion Agency) Japanese dictation free software project. IPA Japanese dictation free software project (April 1997 to March 2000) is aiming at building Japanese LVCSR free software/models based on the IPSJ LVCSR speech database(JNAS) and Mainichi newspaper article text corpus. The software repository as the product of the IPA project is available to the public. More than 500 CD-ROMs have been distributed. The performance evaluation was carried out for the simple version, the fast version, and the accurate version in February 2000. The evaluation uses 200 sentence utterances from 46 speakers. The gender-independent HMM models and 20k/60k language models are used for evaluation. The accurate version with the 2000 HMM states and 16 Gaussian mixtures shows 95.9 percent word correct rate.

2.2 Mian Du, Peter von Etter, Mikhail Kopotev, Mikhail Novikov, Natalia Tarbeeva, Roman Yangarber[3] in the paper "Building Support Tools for Russian-Language Information Extraction" proposed that, there is currently a paucity of publicly available NLP tools to support analysis of Russian-language text. This especially concerns higher-level applications, such as Information Extraction. We present work on tools for information extraction from text in Russian in the domain of on-line news. On the lower level we employ the AOT toolkit for natural language processing, which provides modules for morphological analysis and partial syntactic chunking. Since the outputs of both lower-level modules contain unresolved ambiguity, we synthesize the outputs and pass the result into a pre-existing English-language analysis pipeline. We describe how the information extraction system is adapted for multi-lingual support, including extensions to the ontologies and to the pattern matching mechanism. While this is work in progress, we present an end-to-end pipeline for event extraction from Russian-language news.

2.3 Ulrika Olsson[4] in the paper "Development of a voice-driven dictation application" studied about applications combining speech recognition and telephony are becoming more and more common as tools for helping people with such things as timetable information and directory assistance. This master's thesis describes the development of one such application for dictation, embedded in a system, developed by Icepeak AB, with several other voice driven services used over the phone.

In a dialogue with the system the user dictates and edits her/his message. Several commands areavailable to the user, such as removing or adding phrases, getting parts or the whole message read by a

speech synthesizer etc. The result of the dictation is then put in a file, named after the time the message was created and what kind of message it was. A message is of a certain type, called a template. This can be regarded as a kind of lexicon, a phrase collection, from which the phrases used to compose the message are gathered. This approach makes the set of utterances which have to be recognized much smaller and therefore the recognition process is more stable.

The thesis also describes the design of a web interface where the user can view and control the contents of the database: the templates and the phrases belonging to them

2.4 Young, S.[5] in the paper, "A review of large-vocabulary continuous-speech" proposed considerable progress has been made in speech-recognition technology over the last few years and nowhere has this progress been more evident than in the area of large-vocabulary recognition (LVR). Current laboratory systems are capable of transcribing continuous speech from any speaker with average word-error rates between 5% and 10%. If speaker adaptation is allowed, then after 2 or 3 minutes of speech, the error rate will drop well below 5% for most speakers. LVR systems had been limited to dictation applications since the systems were speaker dependent and required words to be spoken with a short pause between them. However, the capability to recognize natural continuous-speech input from any speaker opens up many more applications. As a result, LVR technology appears to be on the brink of widespread deployment across a range of information technology (IT) systems. This article discusses the principles and architecture of current LVR systems and identifies the key issues affecting their future deployment. To illustrate the various points raised, the Cambridge University HTK system is described. This system is a modem design that gives state-of-the-art performance, and it is typical of the current generation of recognition systems.

2.5 J.L. Gauvain, L.F. Lamel, G. Adda, M. Adda-Decker[6] in the paper “The LIMSI Continuous Speech Dictation System” proposed first results of the CEC-sponsored project Sqale (Speech recognition Quality Assessment for Linguistic Engineering). In this project the recognition performance of four different systems in four different languages is assessed. The official word-error scores are presented, and the significant differences between the systems are indicated. Finally, the first results on the analysis of the various word error rates are given.

2.6 L. Lamel , M. Adda-decker , J.L. Gauvain[7] in the paper “Issues in Large Vocabulary, Multilingual Speech Recognition ” report on our activities in multilingual, speakerindependent, large vocabulary continuous speech recognition. The multilingual aspect of this work is of particular importance in Europe, where each country has its own national language. Our existing recognizer for American English and French, has been ported to British English and German. It has been assessed in the context of the LRESQALE project whose objective was to experiment with installing in Europe a multilingual evaluation paradigm for the assessment of large vocabulary, continuous speech recognition systems. The recognizer makes use of phone-based continuous density HMM for acoustic modeling and n-gram statistics estimated on newspaper texts for language modeling. The system has been evaluated on a dictation task with read, newspaper-based corpora, the ARPA Wall Street Journal corpus of American English, the WSJCAM0 corpus of British English, the BREF-Le Monde corpus of French and the PHONDAT-Frankfurter Runds.

CHAPTER-3
OBJECTIVE

In order to complete the problem defined, the the objective of project work to be met are:

- Our main objective is to develop an application for Punjabi Dictation
- Make decision about the use of specific library and software
- Construct grammar for the Punjabi language
- Construct database for the words in Punjabi
- Test the grammar on different libraries and softwares

CHAPTER-4
PROBLEM FORMULATION

4.1 Problem Statement:

In this thesis work, the problem formulation for Gurmukhi Dictation in Natuarl Language Processing is defined. It would lead to the development of the grammar for the Punjab language and Gurmuki font so that the speech to text covergence softwares and systems could convert speech data into Gurmukhi font. As the speech to text convergence is available for the various internationl language, but are not available for punjabi language and Gurmukhi font. The model used for this purpose is HMM(Hidden Markov's MOdel) along with the HTK toolkit which are responsible for speech recognization and storing the data to the database. Further many liberaries are available for defining the new grammar for the different languages, so, we would use one of these libraries for defining grammar for our requirement. The output of the data may be extracted through http or stp.

4.2 Tools

4.2.1 Hardware

- 2 GB RAM
- Mirophone
- Speaker
- Intel core Processor
- 500 GB Hard disk

4.2.2 Software Requirements

- Operating System: Linux(14.04)
- Editor: vim, joe, atom

4.3 Planning

As the literature survey for the project is completed and now the ultimate goal is to develop the grammer for the dictation. The work plan for achieving the goal is is given below:

• Develop hypothesis for the problem

• Research for the different tools for achieving the goal

• Look for the software and hardware requiremnet and collect the required tools

• Do research for the different liberaries for developing the grammer

• Do research for the best Gurmukhi unicode

• Do coding for developing the grammer

• Test the developed program on the different softwares.

REFERENCES

[1] Wikipedia The free encyclopedia “Natural language processing,”

[2] Shikano Kiyohiro and Kawahara Tatsuya N “IPA Japanese Dictation Free Software Project,”IEEE communications letters, ISSN: 1312-4244, Issue No.7, Vol.5, Page No. 2679-2683, 2007.

[3] Mian Du, Peter von Etter, Mikhail Kopotev, Mikhail Novikov, Natalia Tarbeeva, Roman Yangarber “Building Support Tools for Russian-Language Information Extraction,”ISSN: 1563-3459, Issue No.04, Vol.04, Page No. 1686-1689, Feb 2013.

[4]Ulrika Olsson “Development of a voice-driven dictation application,” IJCSI International Journal of Computer Science Issues, ISSN: 1694-0814, Issue No.4, Vol.7, Page No.32-36, July 2007.

[5] JYoung, S. “A review of large-vocabulary continuous-speech ,” IEEE computer networks , ISSN:4244-62521, Issue No.11, Vol.10, Page No. 116-119, 2011.

[6] J.L. Gauvain, L.F. Lamel, G. Adda, M. Adda-Decker “The LIMSI Continuous Speech Dictation System,” IEEE computer networks, ISSN: 978-6128, Issue No.13, Vol.1,2012

[7] L. Lamel , M. Adda-decker , J.L. Gauvain “Issues in Large Vocabulary, Multilingual Speech Recognition ,” International Journal of Advanced Research in Computer Engineering & Technology, ISSN: 2278–1323, Issue No.6, Vol.1, Page No. 7-13 , Aug 2012.